1 A Hard Exam

There is a table courseData(id, name, faculty), listing the students registered to privacy course. The identifier id is used as a pseudonym. It is only allowed to learn the distribution of students over faculties, i.e. execute the query Q1 on the data:

Q1: SELECT COUNT(*) FROM courseData GROUP BY faculty;

After exam, the teacher has created another table examData(id, faculty, grade) listing grades of the students, using pseudonymous identifiers id shared with table courseData. The teacher still does not want to reveal even a pseudonymized table, and only allows to run certain aggregation queries on it:

Q2: SELECT COUNT(*) FROM examData GROUP BY faculty;
Q3: SELECT COUNT(*) FROM examData GROUP BY grade;
Q4: SELECT AVG(grade) FROM examData GROUP BY faculty;

The starting point for this exercise, containing implementation of queries Q1-Q4, can be found in dpLab1_init.py. The underlying data can be found in the file dpLab1_data.py.

1.1 Demonstrate privacy issues

The attacker knows that Alice is a math student registered to the privacy course. He wants to learn the following.

1. Has Alice taken the exam?

2. If she has, what is her grade?

The attacker should come up with the answers without looking at this data, but just by applying queries Q1-Q4 to this data. Does he actually need all these queries?

1.2 Add differential privacy to COUNT queries

Let us first consider COUNT queries. Compute the sensitivities of queries. Enhance the queries with Laplace mechanism, making them \( \epsilon \)-DP for \( \epsilon \in 1.0, 0.5, 0.1 \).

1. What the attacker thinks after seeing a noisy output? 10 noisy outputs? 20? 100? How is this number related to \( \epsilon \)?
2. Does seeing the whole histogram of grade counts help to learn more about the count of e.g. grade B?

3. Estimate the error for one numerical output, e.g. the COUNT of grades B with $\epsilon = 1.0$. What is the probability that the result is the true count $\pm 1$? Which $\epsilon$ we should take to make this error $\leq 0.05$?

1.3 Add differential privacy to AVG queries

To analyze an AVG query, it is easier to assume that the number of rows $n$ in the table examData is fixed. Instead of hiding the fact that the record of Alice is in the table, we want to conceal the grade of Alice. That is, we define the adjacent tables $t$ and $t'$ as those whose $n - 1$ rows are identical, and the $n$-th row is different. Enhance the query with Laplace mechanism, making it $\epsilon$-DP for $\epsilon \in 1.0, 0.5, 0.1$.

1.4 Differential privacy of a system

Assume that the number of rows $n$ is fixed, and adjacent tables are defined as in the previous point. Let us assume that the attacker is allowed to execute each of the queries Q1-Q4 exactly once. How should one add noise to these queries, so that the differential privacy of the entire system would be $\epsilon$?

1.5 Group privacy

The attacker knows that Alice, Bob and Chris always learn together, and their grades are strongly correlated. How to ensure $\epsilon$-differential privacy for a group of 3 people?